Rapid evolutionary diversification of the *flamenco* locus across simulans clade *Drosophila* species

Sarah Signor[1*], Jeffrey Vedanayagam[2], Filip Wierzbicki[3,4], Robert Kofler[3], and Eric C. Lai[2]

*Corresponding author: sarah.signor@ndsu.edu

[1]Biological Sciences, North Dakota State University, Fargo, North Dakota, USA

[2]Developmental Biology Program, Sloan-Kettering Institute, 430 East 67th St, ROC-10, New York, NY 10065, USA

[3]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria

[4]Vienna Graduate School of Population Genetics, Vienna, Austria

Abstract

Effective suppression of transposable elements (TEs) is paramount to maintain genomic integrity and organismal fitness. In *D. melanogaster*, *flamenco* is a master suppressor of TEs, preventing their movement from somatic ovarian support cells to the germline. It is transcribed by Pol II as a long (100s of kb), single-stranded, primary transcript, that is metabolized into Piwi-interacting RNAs (piRNAs) that target active TEs via antisense complementarity. *flamenco* is thought to operate as a trap, owing to its high content of recent horizontally transferred TEs that are enriched in antisense orientation. Using newly-generated long read genome data, which is critical for accurate assembly of repetitive sequences, we find that *flamenco* has undergone radical transformations in sequence content and even copy number across *simulans* clade Drosophilid species. *D. simulans flamenco* has duplicated and diverged, and neither copy exhibits synteny with *D. melanogaster* beyond the core promoter. Moreover, *flamenco* organization is highly variable across *D. simulans* individuals. Next, we find that *D. simulans* and *D. mauritiana flamenco* display signatures of a dual-stranded cluster, with ping-pong signals in the testis and embryo. This is accompanied by increased multicopy elements, consistent with these regions operating as functional dual stranded clusters. Overall, the physical and functional diversity of *flamenco* orthologs is testament to the extremely dynamic consequences of TE arms races on genome organization, not only amongst highly related species, but even amongst individuals.

Introduction

Drosophila gonads exemplify two important fronts in the conflict between transposable elements (TEs) and the host – the germline (which directly generates gametes), and somatic support cells (from which TEs can invade the germline) (1, 2). The strategies by which TEs are suppressed in these settings are distinct (3), but share their utilization of piwi-interacting RNAs (piRNAs). These are ~24-32 nt RNAs that are bound by the PIWI subclass of Argonaute proteins, and guide them and associated cofactors to targets for transcriptional and/or post-transcriptional silencing (4–7).

Mature piRNAs are processed from non-coding piRNA cluster transcripts, which derive from genomic regions that are densely populated with TE sequences (7–9). However, the mechanisms of piRNA biogenesis differ between gonadal cell types. In the germline, piRNA clusters are transcribed from both DNA strands through non-canonical Pol II activity (6, 10–12), which is initiated by chromatin marks rather than specific core promoter motifs. Moreover, co-transcriptional processes such as splicing and polyadenylation are suppressed within dual strand piRNA clusters (13, 14). On the other hand, in ovarian somatic support cells, piRNA clusters are transcribed from a typical promoter as a single stranded transcript, which can be alternatively spliced as with protein-coding mRNAs (15–18). These rules derive in large part from the study of model piRNA clusters (i.e. the germline *42AB* and somatic *flamenco* piRNA clusters). For both types, their capacity to repress invading transposable elements is thought to result from random integration of new transposons into the cluster. As such, piRNA clusters are adaptive loci that play central roles in the conflict between hosts and TEs.

The location and activity of germline piRNA clusters are stochastic and evolutionarily dynamic, as there are many copies of TE families in different locations that may produce piRNAs (9, 19). By contrast, somatic piRNA clusters are not redundant and a single insertion of a TE into a somatic piRNA cluster should be sufficient to prevent that TE from further transposition (18, 20). Thus, *flamenco* should contain only one copy per TE family (18), which is true in the *flamenco* locus of *D. melanogaster* (18). *flamenco* is also the only piRNA cluster which produces a phenotypic effect when altered, as germline clusters can be deleted with no consequences.

flamenco has been a favored model for understanding the piRNA pathway since the discovery of piRNA mediated silencing of transposable elements (6). *flamenco* spans >180 kb of repetitive sequences located in *β*-heterochromatin of the X chromosome (21). Of note, *flamenco* was initially identifed, prior to the formal recognition of piRNAs, via transposon insertions that de-repress *gypsy*, *ZAM*, and *Idefix* class elements (21–25). These mutant alleles disrupt the *flamenco* promoter, and consequently abrogate transcription and piRNA production from this locus. By contrast, the recent deletion of multiple model germline piRNA clusters, which eliminate the biogenesis of a bulk of cognate piRNAs, did not de-repress their cognate TEs (9). Thus, the analysis of *flamenco* evolution is presumably more consequential for TE dynamics. Analysis of *flamenco* in various strains of *D. melanogaster* supports that this locus traps horizontally derived TEs to achieve silencing of newly invaded TEs (18). The *flamenco* locus exhibits synteny across the *D. melanogaster* sub-group (26); however, the sequence composition of *flamenco* outside *D. melanogaster* has not been well-characterized (27).

In this study, we compare the *flamenco* locus across 10 strains of simulans-clade species, namely *D. simulans*, *D. mauritiana*, and *D. sechellia*. Analysis of piRNAs from ovaries of five genotypes of *D. simulans* found that *flamenco* is duplicated in *D. simulans*. This duplication is old enough that there is no sequence synteny across copies, even though their core promoter regions and the adjacent *dip1* gene are conserved. *flamenco* has also been colonized by abundant (>40) copies of *R1*, a TE that was thought to insert only at ribosomal genes, and to evolve at the same rate as nuclear genes [21]. Furthermore, between different genotypes, up to 63% of TE insertions are not shared within any given copy of *flamenco*. Despite this, several full length TEs are shared between all genotypes in a similar sequence context. This incredible diversity at the *flamenco* locus, even within a single species, suggests there may be considerable variation in its ability to suppress transposable elements across individuals.

Cross-species comparisons further indicate that functions of *flamenco* have diversified. Data from *D. sechellia* and *D. melanogaster* conform with the current understanding of *flamenco* as a uni-strand cluster. However, we find evidence that *D. simulans* and *D. mauritiana flamenco* can act as a dual strand cluster in testis and embryos, yielding piRNAs from both strands with a ping pong signal. Overall, we infer that the rapid evolution of *flamenco* alleles across individuals and species reflects highly adaptive functions and dynamic biogenesis capacities.

Materials and Methods

Fly strains

The four *D. simulans* lines *SZ232*, *SZ45*, *SZ244*, and *SZ129* were collected in California from the Zuma Organic Orchard in Los Angeles, CA on two consecutive weekends of February 2012 [57–61]. *LNP-15-062* was collected in Zambia at the Luwangwa National Park by D. Matute and provided to us by J. Saltz (J. Saltz pers. comm., [41,53]). *MD251*, *MD242*, *NS137*, and *NS40* were collected in Madagascar and Kenya (respectively) and are described in [50]. The *D. simulans* strain *wxD^1* was originally collected by M. Green, likely in California, but its provenance has been lost (pers. comm. Jerry Coyne). *D. mauritiana (w12)* and *D. sechellia (Rob3c/Tucson 14021-0248.25)* are described in [11].

Long read DNA sequencing and assembly

MD242, four SZ lines and *LNP-15-062* were sequenced on a MinION platform at North Dakota State University (Oxford Nanopore Technologies (ONT), Oxford, GB), with base-calling using guppy (v4.4.2). *MD242*, the four SZ lines, and *LNP-15-062* were assembled with Canu (v2.1) [73] and two rounds of polishing with Racon (v1.4.3) [67]. The CA strains were additionally polished with short reads using Pilon (v1.23) [68](SRR3585779, SRR3585440, SRR3585480, SRR3585391) [60]. The first *wxD*$^{1\text{-}1}$ assembly is described here [12]. *MD251*, *NS137*, *NS40* and *wxD*$^{1\text{-}2}$ were sequenced on a MinION platform by B. Kim at Stanford University. They were assembled with Flye [29], and polished with a round of Medaka followed by a round of pilon [68]. Following this contaminants were removed with blobtools (https://zenodo.org/record/845347, [30]), soft masked with RepeatModeler and Repeatmasker [22,64], then aligned to the *wxD*^{1} as a reference with Progressive Cactus [3]. The assemblies were finished with reference based scaffolding using Ragout [28]. *D. mauritiana* and *D. sechellia* were sequenced with PacBio and assembled with FALCON using default parameters (https://github.com/PacificBiosciences/FALCON)[11].The *D. melanogaster* assembly is described here (47). A summary of the assembly statistics is available in Supplementary Table 1. The quality of cluster assembly was evaluated using CUSCO as described in (19, 48) (Supplementary File 1).

Short read sequencing and mapping

Short read sequencing was performed by Beijing Genomics Institute (BGI) on approximately 50 dissected ovaries from adult female flies (*SZ45*, *SZ129*, *SZ232*, *SZ244*, *LNP-15-062*). Short read libraries from 0-2 hour embryos were prepared from *D. melanogaster*, *wxD*$^{1\text{-}2}$, *D. sechellia*, and *D. mauritiana* (SRAXXX) (49). Small RNA from testis is described in (50, 51). Libraries were filtered for adapter contamination and short reads between 23-29 bp were retained for mapping with fastp (52). The RNA was then mapped to their respective genomes using bowtie (v1.2.3) and the following parameters (-q -v 1 -p 1 -S -a -m 50 --best --strata) (53, 54). The resulting bam files were processed using samtools (55). To obtain unique reads the bam files were filtered for reads with 1 mapping position. To obtain counts files with weighted mapping the bam files were processed using Rsubreads and the featureCounts function (56).

Defining and annotating piRNA clusters

piRNA clusters were defined using proTRAC [52]. piRNA clusters were predicted with a minimum cluster size of 1 kb (option "-clsize 1000"), a P value for minimum read density of 0.07 (option "-pdens 0.07"), a minimum fraction of normalized reads that have 1T (1U) or 10A of 0.33 (option "-1Tor10A 0.33") and rejecting loci if the top 1% of reads account for more than 90% of the normalized piRNA cluster read counts (option "-distr 1-90"), and a minimal fraction of hits on the main strand of 0.25 (option "-clstrand 0.25"). Note that this ties the piRNA clusters to their function such that participation in the ping pong pathway can be inferred from these patterns. Clusters were annotated using RepeatMasker (v. 4.0.7) and the TE libraries described in Chakraborty et al. (2019) [12,64]. The position of *flamenco* was also evaluated based off of the position of the putative promoter, the *dip1* gene, and the enrichment of *gypsy* elements [24]. Fragmented annotations were merged to form TE copies with onecodetofindthemall [5]. Fragmented annotations were also manually curated, particularly because TEs not present in the reference library often have their LTRs and internal sequences classified as different elements.

Aligning the flamenco promoter region

The region around the *flamenco* promotor was extracted from each genotype and species with bedtools getfasta (61). Sequences were aligned with clustal-omega and converted to nexus format (62). Trees were built using a GTR substitution model and gamma distributed rate variation across sites (63). The markov chain monte carlo chains were run until the standard deviation of split frequencies was below .01, around one million generations. The consensus trees were generated using sumt conformat=simple. The resulting trees were displayed with the R package ape (64).

Detecting ping pong signals in the small RNA data

Ping pong signals were detected using pingpongpro [66]. This program detects the presence of RNA molecules that are offset by 10 nt, such that stacks of piRNA overlap by the first 10 nt from the 5' end. These stacks are a hallmark of piRNA mediated transposon silencing. The algorithm also takes into account local coverage and the presence of an adenine at the 10th position. The output includes a z-score between 0 and 1, the higher the z-score the more differentiated the ping pong stacks are from random local stacks.

Results

flamenco in the D. simulans clade

We identified *D. simulans flamenco* from several lines of evidence: piRNA cluster calls from proTRAC, its location adjacent to divergently transcribed *dip1*, the existence of conserved core *flamenco* promoter sequences, and enrichment of *gypsy* elements (Figure 1A-D); Supplementary Table 2). The *flamenco* locus is at least 376 kb in *D. simulans*. This is an expansion compared with *D. melanogaster*, where *flamenco* is only 156 kb (*Canton-S*). In *D. sechellia flamenco* is 363 kb, however in *D. mauritiana* the locus has expanded to at least 840 kb (Supplementary Table 2). This is a large expansion, and it is possible that the entire region does not act as the *flamenco* locus. However, evidence that is does include uniquely mapping piRNAs are found throughout the region and *gypsy* enrichment is consistent with a *flamenco*-like locus

(Supplementary Figure 1). There are no protein coding genes within the region, and while the neighboring genes on the downstream side of *flamenco* in *D. melanogaster* have moved in *D. mauritiana* (*CG40813*- *CG41562* at 21.5 MB), the following group of genes beginning with *CG14621* is present and flanks *flamenco* as it is annotated. Thus in *D. melanogaster* the borders of *flamenco* are flanked by *dip1* upstream and *CG40813* downstream, while in *D. mauritiana* they are *dip1* upstream and *CG14621* downstream. Between all species the *flamenco* promoter and surrounding region, including the *dip1* gene, are alignable and conserved (Figure 1E).

Structure of the flamenco locus

Structure of the flamenco locus

D. melanogaster flamenco bears a characteristic structure, in which the majority of TEs are *gypsy*-class elements in the antisense orientation (79% antisense orientation, 85% of which are *gypsy* elements) (Figure 1D; Supplementary Table 3). This is true in both the *iso-1* and *Canton-S* strains. In *D. simulans*, *flamenco* has been colonized by large expansions of *R1* transposable element repeats such that on average the percent of antisense TEs is only 50% and the percent of the locus comprised of LTR elements is 55%. However, 76% of antisense insertions are LTR insertions, thus the underlying *flamenco* structure is apparent when the *R1* insertions are disregarded (Figure 1D). In *D. mauritiana flamenco* is 71% antisense, and of those antisense elements it is 85% LTRs. Likewise in *D. sechellia* 78% of elements are antisense, and of those 81% are LTRs. *flamenco* retains the overall structure of a canonical *D. melanogaster*-like *flamenco* locus in all of these species, however in *D. simulans* the nature of the locus is somewhat altered by the abundant *R1* insertions (Figure 1D).

flamenco is duplicated in D. simulans

In *D. simulans*, we unexpectedly observed that *flamenco* is duplicated on the X chromosome; the duplication was confirmed with PCR and a restriction digest (Supplementary Table 4). These duplications are associated with a conserved copy of the putative *flamenco* enhancer as well as copies of the *dip1* gene located proximal to *flamenco* in *D. melanogaster* (Figure 1C, 2A). While it is unclear which copy is orthologous to *D. melanogaster flamenco*, all *D. simulans* lines bear one copy that aligns across genotypes. We refer to this copy as *D. simulans flamenco*, and the other copies as duplicates. Otherwise, *flamenco* duplicates do not align with one another and lack synteny amongst their resident TEs. Possible evolutionary scenarios are that the *flamenco* duplication occurred early in the *simulans* lineage, that the clustered evolved very rapidly, or that the duplication encompassed only the promoter region and was subsequently colonized by TEs (Figure 1C, 2A).

The *flamenco* duplicate is absent in the *D. simulans* reference strain, w^{501}, but present in wxD^{1}, suggesting it was polymorphic or absent between the collection of these strains (or was not assembled). The duplicate retains the structure of *flamenco*, with an average of 67% of TEs in the antisense orientation in the duplication of *flamenco*, and 91% of the TEs in the antisense orientation are LTRs. The duplicate of *flamenco* is less impacted by *R1*, with some genotypes having as few as 8 *R1* insertions (Figure 2C).

R1 LINE elements at the flamenco locus

R1 elements are well-known to insert into rDNA genes, are transmitted vertically, and evolve similarly as the genome background rate [21]. They have also been found outside of rDNA genes, but only as fragments. However, as mentioned, *R1* elements are abundant within *flamenco* loci in the *simulans* clade. Outside of *flamenco*, *R1* elements in *D. simulans* are distributed according to expectation, with full length elements occurring only within rDNA (Supplementary

File 6). Within *flamenco*, most copies of *R1* occur as tandem duplicates, creating large islands of fragmented *R1* copies (Figure 2A). They are on average 3.7% diverged from the reference R1 from *D. simulans*. Across individual *D. simulans* genomes, ~99 kb of *flamenco* loci consists of *R1* elements, fully 26% of their average total length. *SZ45, LNP-15-062, NS40, MD251*, and *MD242* contain 4-7 full length copies of *R1* in the sense orientation, even though all but *SZ45* bear fragmented *R1* copies on the antisense strand. (The *SZ45 flamenco* assembly is incomplete). As the antisense *R1* copies are expected to suppress *R1* transposition, *flamenco* may not suppress these elements effectively.

In *D. mauritiana*, *flamenco* harbors abundant fragments or copies of *R1* (19 on the reverse strand and 20 on the forward strand), and only one large island of *R1* elements. In total, *D. mauritiana* contains 84 kb of *R1* sequence within *flamenco*. In *D. mauritiana* there are 8 full length copies of *R1* at the *flamenco* locus, 7 in antisense, which are not obviously due to a segmental or local duplication. Finally, we find that *D. sechellia flamenco* lacks full length copies of *R1*, and it contains only 18 KB of *R1* sequence (16 fragments on the reverse strand). Yet, all the copies are on the sense strand, which would not produce fragments that can suppress R1 TEs. Essentially the antisense copies of *R1* in *D. mauritiana* should be suppressing the TE, but we see multiple full length antisense insertions, and *D. sechellia* has no antisense copies, but we see no evidence for recent *R1* insertions. From this it would appear that whatever is controlling the transposition of *R1* lies outside of *flamenco*.

The presence of long sense-strand *R1* elements within *flamenco* is a departure from expectation [21,72]. There is no evidence of an rDNA gene within the *flamenco* locus that would explain the insertion of *R1* elements there, nor is there precedence for the large expansion of *R1* fragments within the locus. Furthermore, the suppression of *R1* transposition does not appear to be controlled by *flamenco*.

piRNA production from R1

On average *R1* elements within the *flamenco* locus of *D. simulans* produce more piRNA than any other TE within *flamenco* (Supplementary Table 6). *R1* reads mapping to the forward strand constitute an average of 51% of the total piRNAs within the *flamenco* locus from the maternal fraction, ovary, and testis using weighted mapping. The only exception is the ovarian sample from *SZ232* which is a large outlier at only 5%. However reads mapping to the reverse strand account for an average of 84% of the piRNA being produced from the strand in every genotype and tissue – maternal fraction, testis, or ovary. If unique mapping is considered instead of weighted these percentages are reduced by approximately 20%, which is to be expected given that *R1* is present in many repeated copies. Production of piRNA from the reverse strand seems to be correlated with elements inserted in the sense orientation, of which the vast majority are *R1* elements in *D. simulans* (Supplementary Figure 2). The production of large quantities of piRNA cognate to the *R1* element is seemingly pointless – if *R1* only inserts at rDNA genes and are vertically transmitted there is little reason to be producing the majority of piRNA in response to this element.

In *D. sechellia* there are very few piRNA produced from *flamenco* in these tissues, and there are no full length copies of *R1*. Likewise overall weighted piRNA production from *R1* elements on either strand is 2.8-5.9% of the total mapping piRNA. In contrast in *D. mauritiana* there are full length *R1* elements and abundant piRNA production in the maternal fraction and testis. In *D. mauritiana* an average of 28% of piRNAs mapping to the forward strand of *flamenco* are arising from *R1*, and 33% from the reverse strand. In *D. mauritiana R1* elements make up a smaller proportion of the total elements in the sense orientation (24%), versus *D. simulans* (55%).

Conservation of flamenco

The *dip1* gene and promoter region adjacent to each copy of *flamenco* are very conserved both within and between copies of *flamenco* (Figure 2). The phylogenetic tree of the area suggests that we are correct in labeling the two copies as the original *flamenco* locus and the duplicate (Figure 2). The original *flamenco* locus is more diverged amongst copies while the duplicate clusters closely together with short branch lengths (Figure 2). They are also conserved and alignable between *D. melanogaster*, *D. sechellia*, *D. mauritiana*, and *D. simulans* (Figure 1). However, the same is not true of the *flamenco* locus itself. Approximately 3 kb from the promoter *flamenco* diverges amongst genotypes and species and is no longer alignable by traditional sequence-based algorithms, as the TEs are essentially a presence/absence that spans multiple kb. There is no conservation of *flamenco* between *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana* (Figure 3). However, within the *simulans* clade many of the same TEs occupy the locus, suggesting that they are the current genomic invaders in each of these species (Figure 3).

In *D. simulans* the majority of full length TEs are singletons – 52% in *flamenco* and 64% in the duplicate. Copies that are full length in one genotype but fragmented in others are counted as shared, not singletons. Almost half of these singletons in the duplicate are due to a single genotype with a unique section of sequence, in this case *MD251*. Likewise a third of the singleton insertions in the duplicate are due to an *NS40* specific region of *flamenco*. Regardless of these concentrations of singletons in single genotypes, it is the single largest category of transposable element insertions, followed by fixed insertions. Thus even within a single population there is considerable diversity at the *flamenco* locus, and subsequently diversity in the ability to suppress transposable elements. For example, *gypsy-29* is present in three genotypes either in *flamenco* or the duplicate, which would suggest that these genotypes are able to suppress this transposable element in the somatic support cells of the ovary while the other genotypes are not. In contrast *gypsy-3* is present in more than one full length copy in *flamenco* and its duplicate it every genotype but one where it is present in a single copy. There are a number of these conserved full length TEs that are present in all or nearly all genotypes, including *Chimpo*, *gypsy-2*, *Tirant*, and *gypsy-4*. In addition, the *INE1* elements adjacent to the promoter are always conserved.

It is notable that any full length TEs are shared across all genotypes, given that *wxD*^{1} was like collected 30-50 years prior to the others, and the collections span continents (Figure 2). Two facts are relevant to this observation: (1) TEs were shown not correlate with geography [32] and (2) *D. simulans* is more diverse within populations than between different populations [38,54,62]. Other explanations are also plausible. Selection could be maintaining these full length TEs, *wxD*^{1} could have had introgression from other lab strains, or a combination of these explanations.

Suppression of TEs by the flamenco locus and the trap model of TE control

In *D. melanogaster*, it was proposed that while germline clusters may have many insertions of a single TE, the somatic 'master regulator' *flamenco* will have a single insertion of each transposon, after which they are silenced and no longer able to transpose [72].

Here, we evaluate the following lines of evidence to determine if they support the trap model of transposable element suppression. (1) How many TEs have antisense oriented multicopy elements within *flamenco*? (2) How many TEs have full length and fragmented insertions, suggesting the older fragments did not suppress the newer insertion? (3) How many *de novo* insertions of TEs in the *flamenco* duplicate of *D. simulans* are also present in the original *flamenco* copy?

<u>How many TEs have antisense oriented multicopy elements within *flamenco*?</u>

Due to the difficulty in classifying degraded elements accurately, for example between multiple classes of *gypsy* element, we will focus here on full length TEs, suggesting recent transposition. In *D. melanogaster* there are 7 full length elements, none of which are present in more than one antisense copy. These elements make up 27% of the *flamenco* locus. Full length copies of five of these elements were also reported previously for other strains of *D. melanogaster* (18)

In *D. sechellia* there are 14 full length TEs within the *flamenco* locus, three of which are present in multiple copies. Two of these, *INE1* and *412*, are likely present due to local duplication. In particular the *INE1* elements flank the promoter, are in the sense orientation, and are conserved between *D. sechellia*, *D. mauritiana*, and *D. simulans*. The only element present in multiple antisense copies is *GTWIN*. Similar to *D. melanogaster* these elements make up 27% of the *flamenco* locus.

D. mauritiana contains 22 full length TEs within the *flamenco* locus. Four of these are present in multiple antisense full length copies – *INE1*, *R1*, *Stalker-4*, and *Cr1a*. While some of the five antisense copies of *R1* likely originated from local duplications – they are in the same general region and tend to be flanked by *gypsy-8*, not all of them show these patterns. Furthermore, as aforementioned, there also are full length sense copies of *R1* suggesting *R1* is not being suppressed by *flamenco*. *gypsy-12* and *gypsy-3* have a second antisense copy within *flamenco* that is just below the cutoff to be considered full length – in *gypsy-3* the second copy is 10% smaller, for *gypsy-12* it is 80% present but missing an LTR. Full length TEs make up 19% of the *flamenco* locus.

In *D. simulans* there are 29 full length TEs present in any of the seven complete *flamenco* assemblies. Eight of these are present in multiple antisense copies within a single genome – *INE1*, *Chimpo*, *copia*, *gypsy-3*, *gypsy-4*, *412*, *Tirant*, and *BEL-unknown*. The two *Tirant* copies are likely a segmental duplication as they flank an *R1* repeat region. In addition, most *INE1* copies are present proximal to the promoter as aforementioned, however in *NS40* a copy is present in antisense at the end of the locus. *Chimpo* is present in three full length copies within *MD242 flamenco*, with no evidence of local duplication. While there are no full length copies of *R1* inserted in antisense, *R1* is present in full length sense copies despite many genomes containing antisense fragments, suggesting *flamenco* is not suppressing *R1*. On average full length TEs constitute 20% of *flamenco* in *D. simulans*.

In the duplicate of *flamenco* in *D. simulans* there are 30 full length TEs present in any one of the five complete *flamenco* duplicate assemblies. However, none of them are multicopy in antisense. However, they are multicopy relative to the original copy of *flamenco*. *gypsy-3*, *BEL-unknown*, *Nomad-1*, *Chimpo*, *gypsy-53A*, *R1*, and *INE1* are all multicopy with respect to the original *flamenco* within a given genome. Some of these may have been inherited at the time of duplication, however are full length in both copies suggesting recent transposition. In the duplicate of *flamenco* full length TEs occupy an average of 17% of the locus. *MD251* is an exception which weights the average, with 28% of the locus, while between 10 and 15% is found for the remaining copies. Thus *D. simulans* and *D. mauritiana* overall do not meet the expectation that *flamenco* will contain a single insertion of any given TE.

How many TEs have full length and fragmented insertions?

Full length elements are younger insertions than fragmented insertions. If a full length element is inserted in *flamenco* and there are fragments in the antisense orientation elsewhere in *flamenco* this indicates that *flamenco* did not successfully suppress the transposition of this element.

In *D. melanogaster* two elements have fragments in antisense and a full length TE – *Doc* and *Stalker-2*. *D. sechellia* has 9 elements that are present as a full length TE and a fragment in antisense (including *412*, *GTWIN*, *mdg-1*, and *nomad*) and 6 that are multicopy that are due to a solo LTR (including *blood*, *297*, and *Stalker-4*). *D. mauritiana* has 21 elements that are present in full length and a fragment in antisense (including *blood*, *412*, *gypsy-10-13*, and *R1*), and four elements that are multicopy due to a solo LTR (*mdg-1*, *Idefix*, and *gypsy-7,10).*

In *D. simulans*, TEs that fit this criteria in *flamenco* include *gypsy-2, gypsy-3*, *gypsy-4, gypsy-5*, *Chimpo*, *412*, *INE1*, *R1*, *Tirant*, and *Zam*. *297* and *Nomad-1* are present in full length copies but only multi-copy in the context of solo LTRs. In the duplicate of *flamenco* in *D. simulans* this includes *gypsy-2*, *gypsy-3*, *gypsy-5*, *297*, *Stalker-4*, and *R1*. For example in *NS40* there are 7 full length copies of *R1* in the sense orientation that likely duplicated in place, as well as 12 partial copies in the antisense orientation. In the *simulans* clade either fragments of TEs are not sufficient to suppress transposable elements or some elements are able to transpose despite the hosts efforts to suppress them.

Is flamenco a trap for TEs entering through horizontal transfer?

High sequence similarity between TEs in different species suggests horizontal transfer [36]. However, because sequence similarity can also exist due to vertical transmission we will use sequence similarity between R1 elements (inserted at rDNA genes) as a baseline for differentiating horizontal versus vertical transfer. There has never been any evidence found for horizontal transfer of *R1* and it is thought to evolve at the same rate as nuclear genes in the *melanogaster* subgroup [21,72]. Of the full length elements present in any genome at *flamenco* 62% of them appear to have originated from horizontal transfer. This is similar to previous estimates for *D. melanogaster* in other studies [72]. Transfer appears to have occurred primarily between *D. melanogaster*, *D. sechellia*, and *D. willistoni*. This includes some known horizontal transfer events such as *Chimpo* and *Chouto* [7], and others which have not been recorded such as *gypys-29* (*D. willistoni*) and the *Max-element* (*D. sechellia*) (Supplemental File 3). The duplicate of *flamenco* is similar, with 53% of full length TEs originating from horizontal transfer. They are many of the same TEs, with a 46% overlap, thus *flamenco* and its duplicate are trapping many of the same TEs. Both *flamenco* and the duplicate the region appears to serve as a trap for TEs originating from horizontal transfer.

In *D. melanogaster* 85% of full length TEs appear to have arisen through horizontal transfer, primarily with *D. yakuba* and *D. sechellia* [72]. In D. sechellia 53% of full length TEs have arisen from horizontal transfer, including some known to have moved by horizontal transfer such as *GTWIN* (*D. melanogaster*/*D. erecta*) [7]. *D. mauritiana* has 68% of its full length TEs showing a closer relationship than expected by vertical descent with TEs from *D. sechellia*, *D. melanogaster*, and *D. simulans*. The hypothesis that *flamenco* serves as a trap for TEs entering the population through horizontal transfer holds throughout the *simulans* clade.

Flamenco piRNA is expressed in the testis and the maternal fraction

Canonically, *flamenco* piRNA is expressed in the somatic follicular cells of the ovary and not in the germline, and also does not produce a ping pong signal [46]. It was not thought to be present in the maternal fraction of piRNAs or other tissues. However, that appears to be variable in different species (Figure 4). We examined single mapping reads in the *flamenco* region from testes and embryos (maternal fraction) in *D. simulans*, *D. mauritiana*, *D. sechellia*, and *D.*

melanogaster. In *D. simulans* and *D. mauritiana flamenco* is expressed bidirectionally in the maternal fraction and the testis, including ping pong signals on both strands (Figure 4). In *D. sechellia*, there is no expression of *flamenco* in either of these tissues. Using weighted mapping in the maternal fraction 63% (*D. mauritiana*) – 36% (*D. simulans*) of the ping pong signatures on the X with a z-score of at least 0.9 are located within *flamenco* (Figure 4). Similar patterns are seen in the testis, with 50% (*D. mauritiana*) to 40% (*D. simulans*) of ping pong signals on the X with a z-score of at least 0.9 being located within *flamenco*. In *D. melanogaster*, there is uni-strand expression in the maternal fraction, but it is limited to the region close to the promoter. In *D. melanogaster* no ping pong signals have a z-score of 0.9, however of those with a z-score of at least 0.8 only 2.3% of those on the X are potentially located within *flamenco*, suggesting that the role of *flamenco* in these tissues has evolved between species.

In the duplicate of *flamenco* in the maternal fraction 18% of the ping pong signals on the X are within the *flamenco* duplicate, while in the testis this is 13%. While overall expression of unique piRNAs is lower, proportionally the locus appears to behave the same in each tissue as the original copy of *flamenco*. In addition, *flamenco* in these species has been colonized by full length TEs thought to be germline TEs such as *blood*, *burdock*, *mdg-3*, *Transpac*, and *Bel* [16,20]. *blood* is also present in *D. melanogaster* in a full length copy while there is no evidence of germline activity for *flamenco* in *D. melanogaster*, though no other putative germline

Silencing of transposable elements

Discussion

The piRNA pathway is the organisms primary mechanism of transposon suppression. While the piRNA pathway is conserved, the regions of the genome that produce piRNA are labile, particularly in double stranded germline piRNA clusters [23]. The necessity of any single cluster for TE suppression in the germline piRNA pathway is unclear, but likely redundant [23]. However, *flamenco* is thought to be the master regulator of the somatic support cells of the ovary, preventing *gypsy* elements from hopping into germline cells [19,42,45,46,48,72]. It is not redundant to other clusters, and insertion of a single element into *flamenco* in *D. melanogaster* is sufficient to initiate silencing. Here we show that the function of *flamenco* appears to have diversified in the *D. simulans* clade, acting in at least some tissues as a germline piRNA cluster.

Dual stranded expression of flamenco

In this work, we showed that piRNAs of the *flamenco* locus in *D.simulans* and *D. mauritiana* are deposited maternally, align to both strands, and exhibit ping-pong signatures. This is in contrast to *D. melanogaster*, where *flamenco* acts as a uni-strand cluster in the soma [40], our data thus suggest that the *flamenco* locus in *D. simulans* and *D. mauritiana* acts as a dual-strand cluster in the germline. In *D. sechellia* the attributes of *flamenco* uncovered in *D. melanogaster* appear to be conserved – no expression in the maternal fraction and the testis and no ping pong signals. Given that *flamenco* is likely a somatic uni-strand cluster in *D. erecta*, we speculate that the conversion into a germline cluster happened in the *simulans* clade [40]. Such a conversion of a cluster between the somatic and the germline piRNA pathway is not unprecedented. For example, a single insertion of a reporter transgene triggered the conversion of the uni-stranded cluster *20A* in *D. melanogaster* into a dual-strand cluster [37].

The role of *flamenco* in *D. simulans* and *D. mauritiana* as the master regulator of piRNA in somatic support cells may still well be true – the promoter region of the *flamenco* cluster is conserved between species and between copies of *flamenco* within species. This suggests that in at least some contexts (or all) the cluster is still serving as a unistrand cluster transcribed from a traditional RNA Pol II site [24]. However it has acquired additional roles, producing dual strand piRNA and ping pong signals, in these two species, in at least the germline and testis. However, in *D. simulans*, the majority of these reverse stranded piRNAs are emerging from the *R1* insertions within *flamenco*. There is no evidence at present that *R1* has undergone an expansion in function in *D. simulans*, thus it is unclear what, if any, functional impact the reverse stranded piRNAs have at the *flamenco* locus.

Duplication of flamenco in D. simulans

In *D. simulans*, flamenco is present in 2-3 genomic copies, and this duplication is present in all sequenced *D. simulans* lines. The *dip1* gene and putative *flamenco* promoter flanking the duplication also has a high similarity in all sequenced lines (Fig. 2B). This raises the possibility that the duplication of *flamenco* in *D. simulans* was positively selected. Such a duplication may be beneficial as it increases the ability of an organism to rapidly silence TEs. Individuals with large piRNA clusters (or duplicated ones) will accumulate fewer deleterious TE insertions than individuals with small clusters (or non-duplicated ones), and duplicated clusters may therefore confer a selective advantage [27].

Rapid evolution of piRNA clusters

A previous work showed that dual- and uni-strand clusters evolve rapidly in *Drosophila* [70]. In agreement with this work we also found that the *flamenco*-locus is rapidly evolving between and within species (Fig. 1C, 3B). A major open question remains whether this rapid turnover is driven by selection (positive or negative) or an outcome of neutral processes (eg. high TE activity or insertion bias of TEs). These rapid evolutionary changes at the *flamenco* locus, a piRNA master locus, suggest that there is a constant turnover in patterns of piRNA biogenesis that potentially leads to changes in the level of transposition control between individuals in a population.

Funding

This work was supported by the National Science Foundation Established Program to Stimulate Competitive Research (NSF-EPSCoR-1826834 and NSF-EPSCoR-2032756) to SS and the Austrian Science Fund FWF (https://www.fwf.ac.at/;) grant P35093 to RK. J.V. was supported by a Pathway to Independence award from the National Institute of General Medical Sciences (K99-GM137077). E.C.L. was supported by the National Institute of General Medical Sciences (R01-GM083300) and National Institutes of Health MSK Core Grant (P30-CA008748).

Competing interests

We declare that we have no competing interests.

Acknowledgements

S.S. would like to thank C. & F. & S. Emery for insightful commentary on the manuscript.

Authors' contributions

S.S. conceived the study, performed bioinformatics and drafted portions of the manuscript. FW and RK performed bioinformatics and drafted portions of the manuscript. JV contributed data and bioinformatic analysis. EL drafted portions of the manuscript and provided data.

Availability of data and materials

All data has been made available in the following repositories:

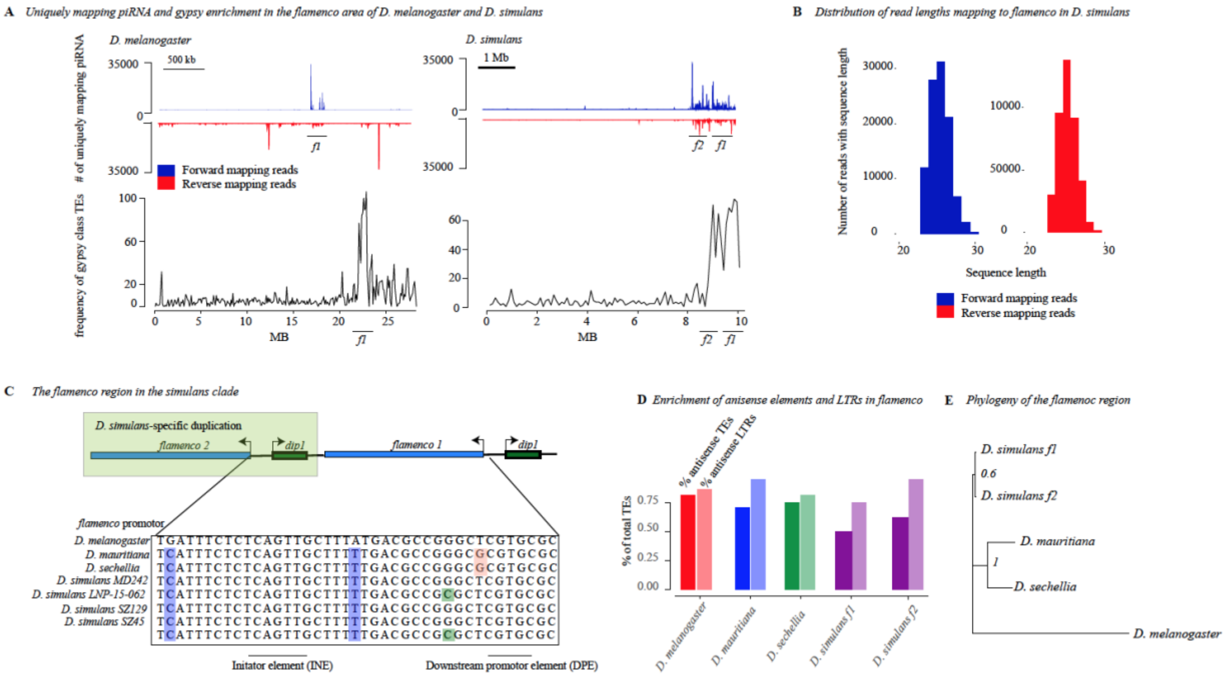

Figure 1. A) Unique piRNA from the ovary and *gypsy* enrichment around *flamenco* and its duplicate in *D. simulans* and *D. melanogaster*. piRNA mapping to the entire contig that contains *flamenco* is shown for both species. The top of the panel shows piRNA mapping to f*lamenco* and is split by antisense (blue) and sense (red) piRNA . The bottom panel shows the frequency of *gypsy*-type transposon annotations across the contig containing *flamenco,* counted in 100 kb windows. There is a clear enrichment of *gypsy* in the area of *flamenco* and, in *D. simulans*, its duplicate compared to the rest of the contig. B) The distribution of read size for small RNA mapping to *flamenco*. The peak is at approximately 26 bp, within the expected range for piRNA. C) The duplication of *flamenco* in the *D. simulans*. Both copies are flanked by the *dip1* gene and copies of the putative *flamenco* promoter. Polymorphisms within the promoter that are shared within the *simulans* clade are shown in blue, *D. simulans* specific polymorphisms are shown in green. The region around the promoter is very conserved across species. D) The percent of TEs in *flamenco* in each species which are in the antisense orientation (first bar) and the percent of TEs in the antisense orientation that are also LTR class elements (second bar). E) A phylogenetic tree of the *dip1* and *flamenco* enhancer region for *D. melanogaster* and the *simulans* clade. This region is conserved and alignable between all species. The tree was generated with Mr. Bayes [51].

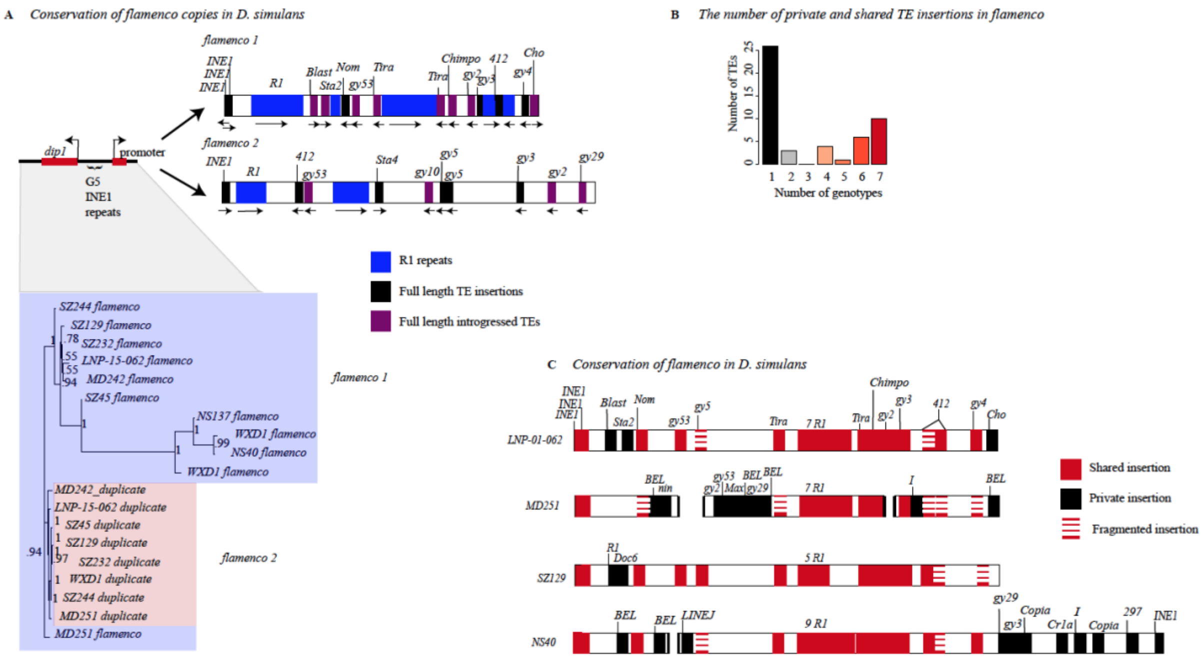

Figure 2. A) Divergence between copies of flamenco. Proximal is a phylogenetic tree of *dip1* and the *flamenco* promoter region from each genome. In between *dip1* and the promoter are a series of *G5/INE1* repeats that are found in every genome. Overall this region is fairly conserved, with the duplicate copies all grouping together with short branch lengths (shown in pink). The original copy of *flamenco* is more diverse with some outliers (shown in light blue) but there is good branch support for all the deep branches of the tree. Distal is a representation of *flamenco* and its duplicate. R1 repeat regions are shown in blue. Full length transposable elements are labeled. There is no synteny conservation between *flamenco* and its duplicate. B) The proportion of insertions that are shared by one through seven genotypes (genotypes with complete *flamenco* assemblies). C) Divergence of flamenco within *D. simulans*. Labeled TEs correspond to elements which are present in a full length copy in at least one genome. If they are shared

between genomes they are labeled in red, if they are unique they are black. If they are full length in one genome and degraded in other genomes they are represented by stacked dashes. If they are present in the majority of genomes but missing in one, it is represented as a missing that TE, which is agnostic to whether it is a deletion or the element was never present

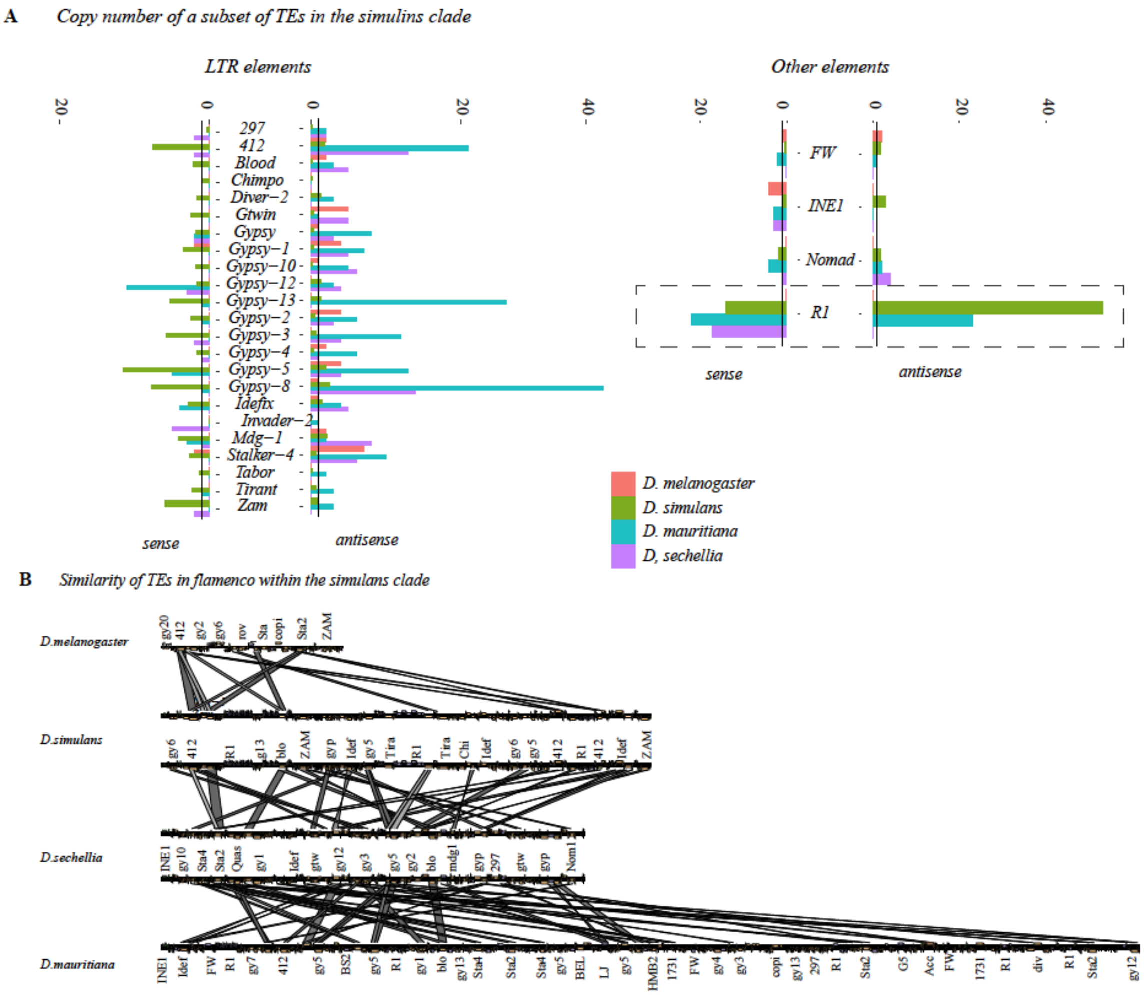

Figure 3) A. Copy number of a subset of transposable elements at *flamenco*. Solo LTRs are indicated by in a lighter shade at the top of the bar. The black line on each bar graph indicates a copy number of one. Values for *D. simulans* are the average for all genotypes with a complete *flamenco* assembly. Note that in *D. melanogaster* (green) most TEs have a low copy number. The expansion of *R1* elements in the *simulans* clade is clearly indicated on the right hand panel with a dotted box. Many elements within *flamenco* are multicopy in the *simulans* clade. While some of this is likely due to local duplications it is clearly a different pattern than *D. melanogaster*. Enrichment of LTR elements on the antisense strand is clear for all species. **B.** Alignment of *flamenco* in *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana*. There is no conserved synteny between species but there are clearly shared TEs, particularly within the *simulans* clade. The expansion of *D. mauritiana* compared to the other species is apparent.

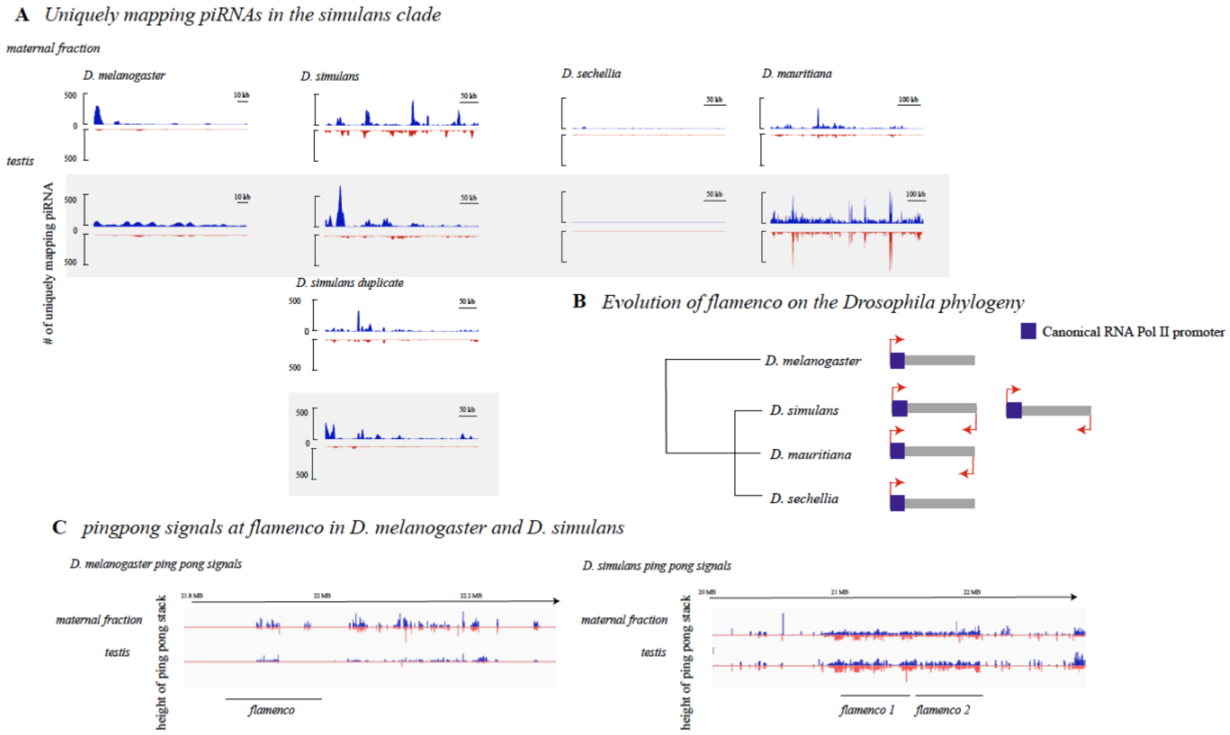

Figure 4) A. Expression of single mapping piRNAs in the maternal fraction and testis (gray) of *D. melanogaster* and the *simulans* clade. Antisense mapping reads are shown in blue, sense in

red. Libraries are RPM normalized and scaled across library type. *D. sechellia* has no expression of *flamenco* in the maternal fraction or the testis. *D. melanogaster* has low expression in the maternal fraction and very little ping pong activity. *D. simulans* and *D. mauritiana* show dual stranded expression in the testis and maternal fraction. **B.** A schematic of the evolution of *flamenco* and its mode expression in the *simulans* and *melanogaster* clade. **C.** *D. simulans* and *D. mauritiana* (Supplementary File) have ping pong singles at *flamenco* in the testis and maternal fraction, while *D. melanogaster* does not.

1. C. Duc, *et al.*, Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. *Genome Biol* 20, 127 (2019).

2. B. Barckmann, *et al.*, The somatic piRNA pathway controls germline transposition over generations. *Nucleic Acids Res* 46, gky761- (2018).

3. C. D. Malone, *et al.*, Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. *Cell* 137, 522–535 (2009).

4. L. S. Gunawardane, *et al.*, A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5' End Formation in Drosophila. *Science* 315, 1587–1590 (2007).

5. S. H. Wang, S. C. R. Elgin, Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. *Proc National Acad Sci* 108, 21164–21169 (2011).

6. J. Brennecke, *et al.*, Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell* 128, 1089–1103 (2007).

7. A. A. Aravin, *et al.*, The Small RNA Profile during Drosophila melanogaster Development. *Developmental Cell* 5, 337–350 (2003).

8. G. Chirn, *et al.*, Conserved piRNA Expression from a Distinct Set of piRNA Cluster Loci in Eutherian Mammals. *Plos Genet* 11, e1005652 (2015).

9. D. Gebert, *et al.*, Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. *Mol Cell* 81, 3965-3978.e5 (2021).

10. P. R. Andersen, L. Tirian, M. Vunjak, J. Brennecke, A heterochromatin-dependent transcription machinery drives piRNA expression. *Nature* 549, 54–59 (2017).

11. C. Klattenhoff, *et al.*, The Drosophila HP1 Homolog Rhino Is Required for Transposon Silencing and piRNA Production by Dual-Strand Clusters. *Cell* 138, 1137–1149 (2009).

12. F. Mohn, G. Sienski, D. Handler, J. Brennecke, The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. *Cell* 157, 1364–1379 (2014).

13. Y.-C. A. Chen, *et al.*, Cutoff Suppresses RNA Polymerase II Termination to Ensure Expression of piRNA Precursors. *Mol Cell* 63, 97–109 (2016).

14. F. Mohn, G. Sienski, D. Handler, J. Brennecke, The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. *Cell* 157, 1364–1379 (2014).

15. C. Goriaux, S. Desset, Y. Renaud, C. Vaury, E. Brasset, Transcriptional properties and splicing of the flamencopi RNAcluster. *EMBO reports* 15, 411–418 (2014).

16. G. Sienski, D. Dönertas, J. Brennecke, Transcriptional Silencing of Transposons by Piwi and Maelstrom and Its Impact on Chromatin State and Gene Expression. *Cell* 151, 964–980 (2012).

17. C. Dennis, E. Brasset, C. Vaury, flam piRNA precursors channel from the nucleus to the cytoplasm in a temporally regulated manner along Drosophila oogenesis. *Mobile DNA* 10, 203–9 (2019).

18. V. Zanni, A. Eymery, M. C. P. of the, 2013, Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *National Acad Sciences* https:/doi.org/10.1073/pnas.1313677110/-/dcsupplemental.

19. F. Wierzbicki, R. Kofler, S. Signor, Evolutionary dynamics of piRNA clusters in Drosophila. *Mol Ecol* (2021) https:/doi.org/10.1111/mec.16311.

20. C. M. Bergman, H. Quesneville, D. Anxolabéhère, M. Ashburner, Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biology* 7, R112-21 (2006).

21. N. Prud'homme, M. Gans, M. Masson, C. Terzian, A. Bucheton, Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. *Genetics* 139, 697–711 (1995).

22. S. U. Song, T. Gerasimova, M. Kurkulos, J. D. Boeke, V. G. Corces, An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. *Genes & development* 8, 2046–2057 (1994).

23. M. Mével-Ninio, A. Pelisson, J. Kinder, A. R. Campos, A. Bucheton, The flamenco Locus Controls the gypsy and ZAM Retroviruses and Is Required for Drosophila Oogenesis. *Genetics* 175, 1615–1624 (2007).

24. A. Pelisson, *et al.*, Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. *The EMBO Journal* 13, 4401–4411 (1995).

25. A. Bucheton, The relationship between the flamenco gene and gypsy in Drosophila: how to tame a retrovirus. *Trends Genet* 11, 349–353 (1995).

26. C. D. Malone, G. J. Hannon, Molecular Evolution of piRNA and Transposon Control Pathways in Drosophila. *Cold Spring Harbor Symposia on Quantitative Biology* 74, 225–234 (2010).

27. A. G. Clark, *et al.*, Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450, 203–218 (2007).

28. D. G. Eickbush, W. C. Lathe, M. P. Francino, T. H. Eickbush, R1 and R2 retrotransposable elements of Drosophila evolve at rates similar to those of nuclear genes. *Genetics* 139, 685–695 (1995).

29. S. A. Signor, F. N. New, S. Nuzhdin, A Large Panel of Drosophila simulans Reveals an Abundance of Common Variants. *Genome Biology and Evolution* 10, 189–206 (2017).

30. S. Signor, S. Nuzhdin, Dynamic changes in gene expression and alternative splicing mediate the response to acute alcohol exposure in Drosophila melanogaster. *Heredity* (2018).

31. S. Signor, Population genomics of Wolbachia and mtDNA in Drosophila simulans from California. *Scientific Reports*, 1–11 (2017).

32. S. A. Signor, M. Abbasi, P. Marjoram, S. V. Nuzhdin, Social effects for locomotion vary between environments in Drosophila melanogaster females. *Evolution* 71, 1765–1775 (2017).

33. S. Signor, Transposable elements in individual genotypes of Drosophila simulans. *Ecology and Evolution* 130, 499–11 (2020).

34. D. R. Matute, J. Gavin-Smyth, G. Liu, Variable post-zygotic isolation in Drosophila melanogaster/D. simulanshybrids. *Journal of Evolutionary Biology* 27, 1691–1705 (2014).

35. D. R. Schrider, J. Ayroles, D. R. Matute, A. D. Kern, Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. sechellia. *PLoS Genetics* 14, e1007341-29 (2018).

36. R. L. Rogers, *et al.*, Landscape of Standing Variation for Tandem Duplications in Drosophila yakuba and Drosophila simulans. *Molecular Biology and Evolution* 31, 1750–1766 (2014).

37. M. Chakraborty, *et al.*, Evolution of genome structure in the Drosophila simulansspecies complex. 139, 1067–63 (2020).

38. , Genome Res.-2017-Koren-gr.215087.116.

39. R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27, 737–746 (2017).

40. B. J. Walker, *et al.*, Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *Plos One* 9, e112963 (2014).

41. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37, 540–546 (2019).

42. D. R. Laetsch, M. L. Blaxter, BlobTools: Interrogation of genome assemblies. *F1000research* 6, 1287 (2017).

43. M. Tarailo-Graovac, N. Chen, Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, 1–14 (2009).

44. J. M. Flynn, *et al.*, RepeatModeler2 for automated genomic discovery of transposable element families. *Proc National Acad Sci* 117, 9451–9457 (2020).

45. J. Armstrong, *et al.*, Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* 587, 246–251 (2020).

46. M. Kolmogorov, *et al.*, Chromosome assembly of large and complex genomes using multiple references. *Genome Res* 28, 1720–1732 (2018).

47. F. Wierzbicki, F. Schwarz, O. Cannalonga, R. Kofler, Generating high quality assemblies for genomic analysis of transposable elements. *Biorxiv*, 2020.03.27.011312 (2020).

48. F. Wierzbicki, F. Schwarz, O. Cannalonga, R. Kofler, Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. *Mol Ecol Resour* 22, 102–121 (2022).

49. Vedanayagam, Jeffrey, "Evolutionary Genomics of piRNA Mediated Transposon Silencing in Drosophila," University of Rochester. (2016).

50. J. Vedanayagam, *et al.*, Endogenous RNAi silences a burgeoning sex chromosome arms race. *Biorxiv*, 2022.08.22.504821 (2022).

51. J. Vedanayagam, C.-J. Lin, E. C. Lai, Rapid evolutionary dynamics of an expanding family of meiotic drive factors and their hpRNA suppressors. *Nat Ecol Evol* 5, 1613–1623 (2021).

52. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Biorxiv*, 274100 (2018).

53. M. J. Axtell, ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA* 19, 740–751 (2013).

54. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25 (2009).

55. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).

56. Y. Liao, G. K. Smyth, W. Shi, The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* 47, gkz114- (2019).

57. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Genome Biology* 34, 3094–3100.

58. D. Rosenkranz, H. Zischler, proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. *Bmc Bioinformatics* 13, 5 (2012).

59. M. Chakraborty, J. J. Emerson, S. J. Macdonald, A. D. Long, Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*, 1–11 (2019).

60. M. Bailly-Bechet, A. Haudry, E. Lerat, “One code to find them all”: a perl tool to conveniently parse RepeatMasker output files. *Mobile Dna-uk* 5, 13 (2014).

61. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).

62. F. Sievers, D. G. Higgins, Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* 27, 135–145 (2018).

63. F. Ronquist, *et al.*, MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* 61, 539–542 (2012).

64. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289–290 (2004).

65. S. Uhrig, H. Klein, PingPongPro: a tool for the detection of piRNA-mediated transposon-silencing in small RNA-Seq data. *Bioinformatics* 35, 335–336 (2018).

66. E. Lerat, *et al.*, Population specific dynamics and selection patterns of transposable element insertions in European natural populations. *Molecular Ecology*, 1–42 (2018).

67. R. S. Singh, Population genetics and evolution of species related to Drosophila melanogaster. *Annual Review of Genetics* 23, 425–453 (1989).

68. H. E. Machado, *et al.*, Comparative population genomics of latitudinal variation in Drosophila simulans and Drosophila melanogaster. *Molecular Ecology* 25, 723–740 (2016).

69. A. Sedghifar, P. Saelao, D. J. Begun, Genomic patterns of geographic differentiation in Drosophila simulans. *Genetics* (2016) https:/doi.org/10.1534/genetics.115.185496.

70. D. A. Petrov, DNA loss and evolution of genome size in Drosophila. *Genetica* 115, 81–91 (2002).

71. E. L. S. Loreto, C. M. A. Carareto, P. Capy, Revisiting horizontal transfer of transposable elements in Drosophila. *Heredity* 100, 545–554 (2008).

72. N. Bargues, E. Lerat, Evolutionary history of LTR-retrotransposons among 20 Drosophila species. *Mobile Dna-uk* 8, 7 (2017).

73. Z. Durdevic, R. S. Pillai, A. Ephrussi, Transposon silencing in the Drosophila female germline is essential for genome stability in progeny embryos. *Life Sci Alliance* 1, e201800179 (2018).

74. B. Czech, J. B. Preall, J. McGinn, G. J. Hannon, A Transcriptome-wide RNAi Screen in the Drosophila Ovary Reveals Factors of the Germline piRNA Pathway. *Mol Cell* 50, 749–761 (2013).

75. G. Coline, E. Théron, E. Brasset, C. Vaury, History of the discovery of a master locus producing piRNAs: the flamenco/COM locus in Drosophila melanogaster. *Frontiers Genetics* 5, 257 (2014).

76. R. Kofler, Dynamics of Transposable Element Invasions with piRNA Clusters. *Molecular Biology and Evolution* 36, 1457–1472 (2019).

77. A. and T. Pélisson, About the origin of retroviruses and the co-evolution of the gypsy retrovirus with the Drosophila flamenco host gene. 29–37 (1997).

78. C. Duc, *et al.*, Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. *Genome Biol* 20, 127 (2019).

79. Y. Luo, P. He, N. Kanrar, K. F. Toth, A. Aravin, Maternally inherited siRNAs initiate piRNA cluster formation https:/doi.org/10.1101/2022.02.08.479612.

80. R. Kofler, piRNA Clusters Need a Minimum Size to Control Transposable Element Invasions. *Genome Biology and Evolution* 12, 736–749 (2020).

81. F. K. Teixeira, *et al.*, piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. *Nature* 552, 268–272 (2017).

82. V. V. Kapitonov, J. Jurka, Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proc National Acad Sci* 100, 6569–6574 (2003).

83. N. D. Singh, D. A. Petrov, Rapid Sequence Turnover at an Intergenic Locus in Drosophila. *Mol Biol Evol* 21, 670–680 (2004).